Available online at: https://ijact.in

Date of Submission	30/03/2019
Date of Acceptance	08/05/2019
Date of Publication	31/05/2019
Page numbers	3126-3134(9 Pages)

Cite This Paper: Namely Driss, Bouzoubaa karim, yousfi abdellah.(2019) A bi-technical analysis for Arabic stop-words detection, 8(5), COMPUSOFT, An International Journal of Advanced Computer Technology. PP. 3103-3111.

ISSN:2320-0790

compusoft
An International Journal of Advanced Computer Technology

A BI-TECHNICAL ANALYSIS FOR ARABIC STOP-WORDS DETECTION

Driss Namly[1], Karim Bouzoubaa[1], Abdellah Yousfi[2]

namly_driss@yahoo.fr, karim.bouzoubaa@emi.ac.ma, yousfi240ma@yahoo.fr
[1]Mohammadia School of Engineers, Mohammed Vth University - Rabat, Morocco.
[2]Faculty of Legal, Economic and Social Sciences-Souissi, Mohammed Vth University-Rabat, Morocco.

Abstract: Stop words are defined as words that frequently appear in texts without carrying any significant information. For the Arabic language, existing works suffer from two main drawbacks (i) the use of only proprietary corpus and (ii) the reliance of only the frequency metric. Our approach for automatic Arabic stop-words detection uses a new metric based on a supervised machine learning process and a vector space representation that can be applied to any corpus, taking into account both domain-independent and domain-dependent stop-words. Conducted experiments to evaluate the proposed approach show a significant improvement reaching 91.85% for the detection rate using the F-measure metric.

Keywords: NLP, Stop-words, Supervised machine learning; Arabic; Information retrieval.

I. INTRODUCTION

Stop-words are general common words of a language, necessary for sentences construction because of their syntactic function, but with no significant semantic added value for documents in terms of information retrieval. Also known as common words, noise words or negative dictionary, stop-words represent 30 to 50% of the textual data size [1]. Therefore, they are not included as indexing terms and their removal leads to higher performances efficiency by reducing useless processing, without affecting retrieval effectiveness [2].

This relevance has given rise to stop-words studies for many languages such as Chinese [3], French [4], Mongolian [5], Arabic [6] or Farsi [7], and their exploitation in numerous text processing fields such as spelling normalization, stemming and stem weighting [4], Text classification [8], document clustering [9] or search engines [10].

All these aforesaid works involve generic stop-word lists called domain-independent stop-words. This kind of stop-words doesn’t depend on the documents being used for their detection. They depend on the concerned language itself. For example, the words “a, the, an, as, that and for” are some English domain-independent stop-words, that can be detected as stop-words using any English corpus. Likewise, the words “من، إلـى، حتـى، ثم، أو، في” (From, to, up, then, or, in) are some Arabic domain-independent stop-words. In this work, we distinguish between domain-independent stop-words and domain-dependent stop-words also named corpus-based stop-words, specialized lexicon or thematic lexicon. These domain-dependent stop-words include words specific to a domain or topic, serving to focus on the knowledge of a particular field such as the words health, sick, healing or pain in the medical domain. These kinds of words are considered as stop-words in the documents handling this particular domain.

Generally, to determine a stop-words list, we use one or a combination of two approaches. The first one is to rely on expert’s judgment to enumerate words following stop-word features. The expert can be a linguist for domain-independent stop-words list or a historian for a history domain-dependent stop-words list. For example, linguists [2] recognize that words belonging to some special syntactic classes such as prepositions, conjunctions, pronouns and adverbs can be considered as stop-words. The second approach is to use corpus statistics to compile stop-words list. To the best of our knowledge, all previous works interested in corpus statistics approach use the frequency of the word in the corpus documents as a weighting scheme to determine the stop-words list. Among these schemes we could name Document Frequency (DF) [11], Term Frequency-Inverse Document Frequency (TF-IDF) [11], Chi-squared statistic [12] and Entropy [3].

In addition, frequency-based stop-word lists depend mainly on the corpus from where the stop-words have been extracted. For example, in some researches, words occurring more than 25000 times [2] are considered as stop-words. This frequency threshold can't be applied to any corpus. Thus, a stop-words list of such kind is completely useless once we change the corpus.

However, stop-words detection using a corpus covering one particular domain such as society, sciences or religion, reveal the presence of some words, different from the usual stop-words but having the same behaviors. Namely they appear in the text very frequently and they are very common in the selected corpus domain. These stop-words are called domain-dependent stop-words. A stop-word in one domain is not necessarily a stop-word in a different domain. For example, the word "معادلة" (Equation) is a stop-word in a collection of articles dealing with mathematics, but certainly not in a collection addressing geography. Also, the word "دولار"(Dollar) is a stop-word in a collection of articles on the subject of finance, but certainly not in a collection discussing philosophy. Henceforth, these domain-dependent stop-words should be detected and processed in the context of information retrieval or text mining applications. However, these domain-dependent stop-words have not been addressed enough in previous works and only few of them are commonly known.

For the Arabic language, in spite of this broad range of frequency-based metrics, there is no commonly accepted stop-words list. Some researchers compile stop-word lists manually based on their experiences or using some existing lists, while other researchers exploit frequency based metrics using their own corpora which might be of any domain or even multi-domain. The outcome of this state is that there is a multitude of stop-words lists with a variable content. This non availability of a standard stop-words list for Arabic language applies to domain-dependent as well as to domain-independent stop-words lists.

The objective of the current paper is to present an innovative method that uses not only frequency-based metrics to detect domain-independent and domain-dependent stop-words, regardless of whether the corpus involves a specific domain or not.

The proposed technique overcomes the previously cited troubles which are:

- The exclusive exploitation of frequency-based metrics;
- The use of a particular corpus;
- The domain-dependent stop-words neglect.

The paper consists of five sections. Related works are reviewed in Section 2. In Section 3, we explain the proposed approach. Section 4 exposes experimental details and results. Section 5 presents the evaluation of the technique. A conclusion is presented in Section 6.

II. Related works

Stop-words detection aims at identifying potential candidate stop-words. In existing works, we present both efforts involving Arabic and non-Arabic language in order to acquire an overall overview of the situation. Let's us note that even if there are many works dealing with stop-words [1, 2, 3, 4, 5, 6, 7, 11, …], we limit this literature review to few of them by retaining the most innovative and representative researches.

A. *Non Arabic stop-words*

Jacques Savoy in [4] aims to propose a general stop-words list and a simple stemming procedure required for French corpora. The author sorted all the word forms appearing in their French corpora according to the frequency of occurrence and extracted the 200 most frequently occurring words. Secondly, he inspected this list to remove all numbers, nouns and adjectives more or less directly related to the main subjects of the underlying collections. The resulting stop-words list were thus a large number of pronouns, articles, prepositions and conjunctions and the suggested stop-words list for French contains 215 words.

Feng Zou et al . [3] aggregated two lists of stop-words using Borda's rule to get a single stop-words list. The two lists are generated by performing a statistical analysis on a corpus of 423 English articles in TIME magazine. The first one is based on a statistical model. The most frequent words and the distribution of word frequencies in different documents statistics are used to refine the stop-words list to get words with stable and high frequency in documents. For this, authors measured the mean of probability and the variance of probability of each word in individual document. The final formula in the statistical model called "the statistical value of a word" (SAT) is the mean divided by the variance and words with highest SAT will be considered as stop-words. The second model is an information model based on the entropy. Therefore, the words with lower entropy are candidate stop-words.

Khalifa Chekima et al . [13] proposed an aggregation technique using three different approaches for an automatic construction of general Malay Stop-words list. The first statistical approach is based on words' frequencies (highest and lowest). The second approach considers words' distribution against documents using variance measure. The third approach computes how informative a word is by using Entropy measure. As a result, a total of 339 Malay stop-words were produced.

B. *Arabic stop-words*

Ibrahim Abu El-Khair [2] generated three stop-words lists. The first one, consisting of 1377 words, is a general stop-words list based on the Arabic language syntactic classes.

The second one is a corpus-based stop-words list, built using words occurring more than 25000 times. This methodology identifies 359 words and the manual check provides a resulting list contained 235 words. The third stop-words list created by combining the general and corpus-based stop-words lists resulted in a list of 1529 words. In this case also, detection metrics is frequency of occurrences which is highly correlated with the used corpus.

A paper by Alajmi et al . [14] presented a statistical approach to extract Arabic stop-words list. Authors generated three lists, the first one is constructed by determining Word Frequency, the second one is based on Mean and variance and the third list is established by calculating the entropy. The three generated lists are aggregated using Borda's Rule to obtain the final list. The extracted list was compared to a general list. The resulting list contains 200 words based on the frequency metric.

Medhat Walaa [6] explored the effect of removing stop-words on a sentiment analysis task. He generated a stop-words list of Egyptian dialect using a list of the most frequent words from an online social network corpus for Egyptian dialect and the MSA. The author filtered by hand a list of the most frequent 200 words to obtain a final list of 100 valid unique words. This list is enriched by adding some frequent prefixes and suffixes. At the end, the final corpus-based list contains 1061 words. The used corpus combines 1261 Facebook comments, 781 tweets and 32 reviews downloaded from the review sites. In addition to the use of a single detection metric which is the frequency of occurrence, data type and size choices are limitations for the acquirement of a commonly accepted stop-words list.

C. Summary

Table I: Works related to stop-words lists detection summary

Authors	*Language*	*Methods*	*Corpus*	*Domain-independent stop-words*	*Resulting list*	*Remark*
Jacques Savoy	French	1. frequency of occurrence	Articles from the French newspaper "Le Monde"	Not covered	215 stop-words	
Feng Zou et al.	Chinese	1. the mean and the variance of each word in individual document 2. an information model based on the entropy	Chinese corpus consisting both of People's Daily news and Xinhua news	Not covered	More than 300 stop-words	aggregated using Borda's rule
Khalifa Chekima et al.	Malay	1. words' frequencies 2. variance measure 3. Entropy measure	A Malay corpus	Not covered	339 stop-words	aggregated
Ibrahim Abu El-Khair	Arabic	1. based on the Arabic language syntactic classes 2. words occurring more than 25000 times	The Arabic Newswire corpus	Covered	1529 stop-words	combination of the two lists
Alajmi et al.	Arabic	1. Word Frequency 2. Mean and variance 3. Entropy	A corpus containing 1002 Documents	Not covered	200 stop-words	aggregated using Borda's Rule
Medhat, Walaa	Egyptian dialect	1. Most frequent words	Online social network corpus	Not covered	1061 stop-words	

Works detailed in the previous section are summarized in Table I.

Regarding the Arabic stop-words, these works undergo the following troubles:

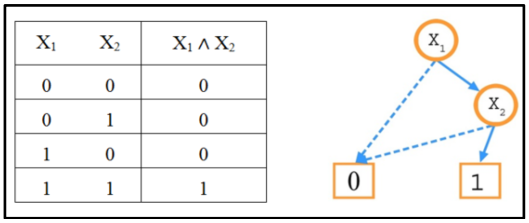
- The extensive use of frequency-based statistical methods, because all the existing works are based exclusively on the frequency or on one of its derived metrics (Mean, Variance and Entropy). So, even if the frequency gives good results, in our point of view, it is not enough to say that a word is a stop-word and there may be another metric, such as words distribution which can provide better results in combination with the frequency;
- Arabic works use their own corpora, which leads to a biased stop-words lists;
- Arabic works deals with domain -independent stop-words (except Abu El-Khair) and don't consider domain-dependent ones.

To overcome these problems, we propose a new approach for an automatic extraction of Arabic stop-words. The idea behind this work emanates from the nonexistence of a commonly accepted Arabic stop-words list. Hence, to avoid previous works weaknesses, the challenge is to propose an approach using other alternative metrics that can be applied to a largely known corpus, taking into account both domain-independent and domain-dependent stop-words.

III. Proposed approach

In this section, we describe our new approach applied to monitor words behaviour to offer an extensive mechanism for stop-word lists detection. To do that, our starting point is stop-words definition.

Generally defined as the most common words in a language, the most frequent terms or common words which would appear to be of little value in helping select documents matching a user need, according to our analysis, we define stop-words as "Terms that occur frequently and are uniformly distributed in most of the documents in a given collection". We represent our definition in a binary decision diagram with its truth table as shown in Figure 1. X_1 represents the frequency whereas X_2 is the distribution. Then, a term can be considered as stop-word if and only if it is a frequent one (X_1 = 1) and uniformly distributed (X_2 = 1) in most of the documents in a given collection.



X_1	X_2	$X_1 \land X_2$
0	0	0
0	1	0
1	0	0
1	1	1

Figure 1: truth table and Decision Tree representation of the definition

Figure 2 shows an example of the four states of the truth table for a specific word in a single document. In the following figure, each sub figure illustrates a text document and the needed word is surrounded by a dark (red) color. Figure 2-a represents a non-frequent and non-distributed word (all occurrences are assembled in the sixth paragraph). Figure 2-b shows a non-frequent word, but distributed in all document paragraphs. Figure 2-c represents a frequent and non-distributed word and finally Figure 2-d represents a frequent and distributed word. Initially, we tried to use a combination of the frequency with the word distribution in the documents to detect stop-word candidates as a transcription of the stop-words definition into a mathematical formalism, but the designed metric gave poor results especially for domain-independent stop-words. This defect can be explained by the fact that the formula used in the distribution modelling is based on the frequency itself.

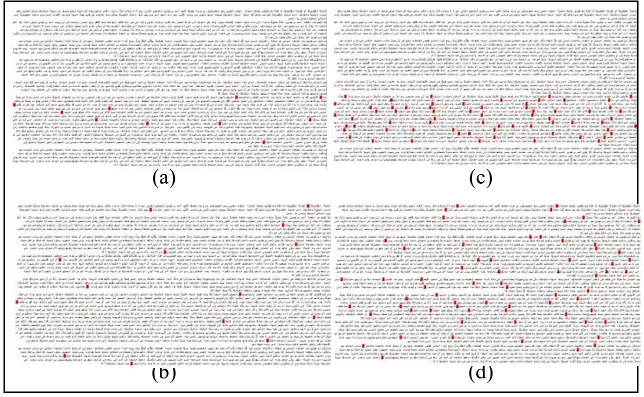


Figure 2: The four states of the truth table in a single document

The second attempt, was to use statistical analyses such as logistic regression with the TF-IDF weighting scheme. This method failed also because the goodness of the model performed using the likelihood ratio test and the pseudo R2 metric indicates a bad fit of the model. From these two attempts rises the third endeavor. On one hand, if a word appears with a high frequency then it is a frequent word. On the other hand, a word that appears in most of the collection documents, in most of the domains and in all or most of the documents segments then it is certainly a well distributed word. This investigation leads us to adopt a bi-technical model for stop-words detection. In a first step, we compute its normalized frequency. This first technique is primarily based on the frequency and brings us the first stop-words list. In a second step we hypothesize that a word that appears in every collection documents, in every domain and every document window might likely be a stop-word because the word exists and is uniformly distributed in all documents. This second technique is based on the distribution represented by the presence of the word in the corpus in the domains and in the articles. The series of distances is used in a customized vector space model which gives us the second stop-words list. The combination of the two lists acquires powerful results and brings a stop-words' list that respects the stop-words definition namely the high frequency with the uniform distribution in most of the documents. To do that, the aggregation method used to group the two lists consists in a logical "and" operation that selects words in common between the two lists.

The first technique involves a supervised machine learning process that uses the normalized words frequency to classify words and identify stop-word candidates. The second technique employs vector space representation to catch stop-word candidates. Subsequently, the two methods are aggregated to get the final stop-words list.

A. The frequency-based technique (Technique A)

To detect stop-words, the first used technique is a supervised machine learning task involving training data analysis to produce a word probability that can be used for new words checking. The objective of the analysis is to discriminate the terminology of the corpus into two classes, stop-words and nonstop-words. In other words, we aim to predict whether a word of the corpus belongs to the stop-words group or not.

As shown in Figure 3, the dataset is split into two parts, a training set and a test set. The training set is a dataset tagged with expected classes which serves to build the supervised learning model in this phase. We also rely on expert's judgment to determine a discrimination threshold that specifies the boundaries between the two classes. This threshold is used to apply the model to the test set data to get the classification results.

So, to build our binary classifier model, we use a set of learning data to estimate the value of a response variable based on the value of an explanatory variable, to identify the two classes. The explanatory variable used in the classification is f = the normalized word frequency in the document space merged and treated as a single document. This probability function is the naive Bayes probabilistic classifier.

The inferred function is written as follows:

$$P(X) = f \qquad (1)$$

Where X is the observation and P(X) is the probability for X.

Thus, to assign observations (words) to each class (stop-words and nonstop-words), a discrimination threshold is imposed on the predicted probabilities P(X).

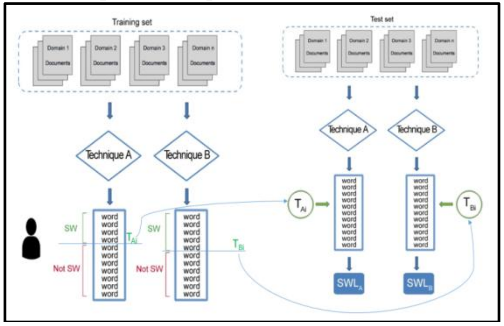


Figure 3: Machine learning workflow

B. *The vector space technique (Technique B)*

The point of interest in this second technique is the adoption of the usual vector space model [15] for representing words presence and not text documents. In the ordinary vector space model, documents are represented as vectors with components corresponding to each term in the dictionary and the value of each component is a weighting metric (number of occurrences, term frequency or term frequency-inverse document frequency) for that word.

In our technique, we want to quantify the similarity between words not documents. That is why our first customization for this model is to inverse the usual vector space model matrix to represent words in a vector space model. The second adaptation is instead of using documents to represent vector dimension, we use words presence in domains, documents and document segments constituted by a fixed width document window of the size 200 words.

In the adjusted model, each word is viewed as a vector of components corresponding to its presence or absence in each window in the documents segment. The advantage of this model against the usual vector space model is the linearity of the dimension, because vectors (words) have the same dimension (the number of documents window) in every level.

So to classify a word, we propose a prototype word that is present in all document segments and the similarity between a word and the prototype is then measured in terms of distances in this vector space by computing the cosine similarity between the two vectors (word vector and prototype vector). Words closest to the prototype (up to a threshold) can be considered as stop-word.

In the classifier model building stage, we estimate the value of a response variable based on the value of a number of other explanatory variables, to identify the two classes, where every explanatory variable is represented by a vector space model. The explanatory variables used in the classification are X_1= the presence rate of the word in the document space (the percentage of documents in which it is present), X_2= the presence rate of the word in a section of the document space (the percentage of pages in which it is present) and X_3= the presence rate of the word in document windows (the percentage of windows in which it is present)

The inferred function is a weighted sum of explanatory variables, which is written as follows:

$$P(X) = w_1{*}\ X_1 + w_2{*}\ X_2 + w_3{*}\ X_3 \qquad (2)$$

To determine the weights w_1, w_2 and w_3, we diversify the explanatory variables weight in order to test the impact of this change on the inferred function by giving more relevance to a specific variable than the others. For that, we tested with the following weight triplets [0.33, 0.33, 0.33], [0.5, 0.3, 0.2], [0.2, 0.3, 0.5], [0.1, 0.3, 0.6] and [0.6, 0.3, 0.1] for [w_1, w_2, w_3]. By doing so, the maximal average variation on the inferred function is 2,33% recorded between the triplets [0.33, 0.33, 0.33] and [0.6, 0.3, 0.1]. Thus, this weights adjustment leads to a non-significant difference in the outcome. As a result, we choose to apply the same weight for all explanatory variables. The final inferred function is an equally weighted sum of explanatory variables since we associate the weight w=1/3 for everyone.

The final output of the model is the equation below:

$$P(X) = 1/3\ {*}(X_1 + X_2 + X_3) \qquad (3)$$

To assign observations (words) to each class (stop-words and nonstop-words), a discrimination threshold is imposed on the predicted probabilities P(X).

IV. Experiment

In this section, we introduce the data set composed by 4027 Arabic Wikipedia documents divided in two sets used in training and testing phases. The training phase consists to handle an initially tagged documents collection to build a reference classification model, whereas, the testing phase involves the use of the model built to classify the untagged documents. Finally, we expose and discuss the obtained results.

A. *Data set*

In this experiment, Arabic Wikipedia[1] is used as a source of documents. This choice is due the nature of Arabic Wikipedia which contains a very large number of documents categorized in numerous domains and expressed with a well written Arabic language. And to evaluate our techniques we use a second corpus (that will be detailed in the next section) gathered from an electronic newspaper.

We extract the data from a dump of Arabic Wikipedia (arwiki-2017-01-11) to build a corpus created from the Wikipedia articles. For that purpose, we keep only textual articles by removing all the other Wikipedia pages (file, image, mediawiki, template, etc.).

Table II: The experimental data set

[1] https://ar.wikipedia.org

	Training documents		Testing documents	
D_0 - General reference	514	50.00%	514	50.00%
D_1 - Religion and belief systems	150	50.00%	150	50.00%
D_2 - People and self	198	50.13%	197	49.87%
D_3 - History and events	140	50.00%	140	50.00%
D_4 - Technology and applied sciences	150	50.00%	150	50.00%
D_5 - Culture and the arts	131	50.19%	130	49.81%
D_6 - Geography and places	269	50.09%	268	49.91%
D_7 - Mathematics and logic	60	50.42%	59	49.58%
D_8 - Natural and physical sciences	160	50.16%	159	49.84%
D_9 - Philosophy and thinking	40	50.63%	39	49.37%
D_{10} - Society and social sciences	205	50.12%	204	49.88%
Total	2017	50.09%	2010	49.91%

As we can see from table II, the corpus is composed of 11 main domains (the domain D_0 for domain-independent stop-words and domains D_1 to D_{10} for domain-dependent stop-words). Each domain is a set of TXT files encoded in UTF-8, named using the ID of the Wikipedia article. However, because of processing limits (time and speed), the final data set contains 4027 Arabic Wikipedia documents of 669.349 words. We consider this size large enough in terms of the richness of the corpus data.

To build our classifier model, we use 50% of the labeled data for training and the remaining 50% for test purposes. Generally, 70-30 ratio is often used, but because of the absence of an initially tagged data and the heaviness of this task we choose the 50-50 ratio.

B. *Experimental steps*

At the beginning of this experiment, we observe that the vocabulary generated using the corpus contains words and their inflected forms in distinct entries. For example, the particle "إِلَى" (To) has more than eighteen inflected forms " إِليهَا، لإِليهِ، لَإِليكُمْ، أَفَإِلى، وَإِلى، أَوَإِليهِ، إِليكُمْ، فَإِليكَ، ...etc" (To him, and to you, to you, and to, is to him, to him, …) and their use gives an incorrect number of dictionary terms and incorrect statistics for these terms. For that, instead of having in the dictionary eighteen terms with wrong statistical values, we add only the lemmatized form which is the word "إِلَى" (To). For that reason, we opt to perform the experiment with a lemmatized version of the collected Wikipedia articles instead of the original ones.

After that, we start the design of the two models. Figure 4 shows the proposed experiment framework which is composed of five steps. We apply this experiment for all domains and every technique.

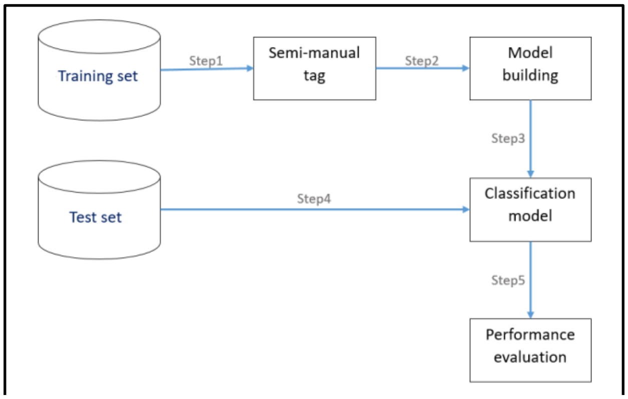


Figure 4: Experimental framework

Step 1: we tag semi-manually the training documents to specify if a word is a domain-independent (DI) stop-word such as " في، من، على، هذا، أو، غير" (non, or, this, on, from, in), domain-dependent (DD) stop-word such as the following words "قياس، حساب، معادلة، استنتاج، تحليل، جبر، زاوية" (measure, count, equation, conclusion, analysis, algebra, angle) for the domain D7 (Mathematics and logic) or not (F) stop-word like " أساسي، تأليف، قناة، شخص، شمس، بساط، مصنف" (basic, author, channel, person, sun, mat, book). Theses tags serve in the third step to build the classifier.

Step 2: The model building stage consists in computing the values of the explanatory variables to get the probabilities P(X). Examples of the obtained results are given in tables III and IV.

Table III presents an extract of the results obtained using the first technique for the domain D1 (Religion and belief systems). We compute, the normalized word frequency for each word. This measure reflects the probability of our binary classifier. For example, the response variable " كتاب" (book) has a probability of 11.33% to be stop-word.

Table III: Inferred function values extract for the domain D_1 using technique A

X		P(X)
في	(in)	96.98
من	(from)	81.72
على	(on)	42.30
الله	(god)	37.61
عن	(about)	18.88
إحسان	(ihsan)	12.99
إسلام	(Islam)	11.93
بين	(between)	11.63
كتاب	(book)	11.33
قرآن	(Quran)	11.18
إسلامي	(Islamic)	9.82
لم	(did not)	9.21
صلاة	(prayer)	6.50
رسول	(messenger)	6.04

Table IV: Inferred function values extract for the domain D_0 using technique B

X		X_1	X_2	X_3	P(X)
في	(in)	94.29	96.64	96.88	95.94
من	(From)	91.43	90.63	94.13	92.06
هذا	(This)	71.43	71.78	77.23	73.48

أو	(or)	57.14	59.73	63.52	60.13
غير	(Non)	51.43	53.27	56.20	53.63
اعتبر	(It is)	31.43	52.58	36.27	40.09
رجل	(man)	11.43	48.11	28.10	29.21
صور	(photo)	2.86	33.33	29.62	21.94
أزمة	(crisis)	2.86	33.33	9.37	15.19
إرسال	(send)	5.71	16.25	13.25	11.74
محلي	(local)	2.86	16.67	9.37	9.63

Table IV presents an extract of the results obtained using the second technique for the domain D_0 (General reference). We compute for each response word, the inferred function composed by the three explanatory variables X_1, X_2 and X_3, where every explanatory variable is represented by a vector space model. Therefore, we get the prediction probability of our binary classifier. For example, the response variable "في" appears in 94.29% of all domains and in 96.64% of the documents and in 96.88% of the documents' windows. As a consequence, the word "في" has the probability of 95.94% to be a stop-word.

Step 3: we use the training data tags to specify the probability thresholds for all domains and for both techniques. The outcome of this step is the threshold value that discriminates the classes. For example, according to expert judgment, the threshold for the domain D_0 using technique B is 52.38%, so that, words having a probability P(X) greater than this value are stop-words.

Step 4: we apply the obtained thresholds on the test set probabilities to predict words' classes. The result of this step is the stop-word lists.

Table V: Confusion matrix

	P	N
T	10036	48
F	2978	566

Step 5: To evaluate the stop-words detection techniques, we compute the precision, recall and F-measure evaluation metrics. However, prior to that, we compute the confusion matrix for each one. A confusion matrix for the domain D_0 and the first technique is shown in Table V.

TableVI presents the evaluation metrics results for the first technique, the second technique and the aggregation of both techniques using the formulas (4, 5 and 6).

$$\text{Precision} = \text{Pr} = \text{TP} / (\text{TP}+\text{FP}) \quad (4)$$

$$\text{Recall} = \text{Rc} = \text{TP} / (\text{TP}+\text{FN}) \quad (5)$$

$$\text{F-measure} = (2 * \text{Pr} * \text{Rc}) / (\text{Pr} + \text{Rc}) \quad (6)$$

Where TP is the number of stop-words predicated as stop-words, TN is the number of stop-words predicated as nonstop-words, FP is the number of nonstop-words predicated as nonstop-words and FN is the number of nonstop-words predicated as stop-words.

For example, using the first technique the F-measure for the domain D_6 reaches 81.36%. We mention that the F-measure records on average 81.63% for the first technique, 85.66% for the second one and 91.85% for the aggregation A ∩ B.

Table VI: Techniques evaluation metrics results

	Technique A	Technique B	A ∩ B
D_0	80.29	82.76	88.78
D_1	79.98	82.26	89.01
D_2	81.87	82.57	94.13
D_3	79.02	81.86	94.87
D_4	85.39	89.62	93.67
D_5	79.96	84.73	89.94
D_6	81.36	86.05	92.82
D_7	84.12	90.26	93.13
D_8	83.74	87.53	88.41
D_9	79.03	85.86	91.07
D_{10}	83.18	88.76	94.54
Avg	81.63	85.66	91.85

V. EVALUATION

To evaluate our techniques, we discuss the obtained metrics results and we confirm this assessment using a second v evaluation metric results show that the combination of both techniques gives the value 91.85% in the average for all domains. From this result we can conclude that our classifier "A ∩ B" discriminates the terminology of the corpus into the three classes (DD, DI and F) with an accepted rate. That means that, the modeled classifier is able to reproduce more that 90% of expert's judgment done in the experimental steps.

To confirm this result, we compare our technique with the Entropy. Our technique is built by aggregating the first technique which is a frequency based and the second one which is a distance based technique, while the entropy is computed using the formula (7) where $P_i(w)$ is the document frequency for the word w in the document i and n is the number of documents in the documents collection.

$$H(w) = \sum_{i=1}^{n} P_i(w)\log[\frac{1}{P_i(w)}] \quad (7)$$

Table VII: evaluation metrics results

	Technique A ∩ B	Entropy
D_0	88.78	82.65
D_1	89.01	80.81
D_2	94.13	83.61
D_3	94.87	81.80
D_4	93.67	86.97
D_5	89.94	80.17
D_6	92.82	81.43
D_7	93.13	86.72
D_8	88.41	86.01
D_9	91.07	81.01
D_{10}	94.54	83.41
Avg	91.85	83.14

Table VIII: aggregate list examples

Domain	Stop-words
D_0: General reference	في، من، على، الى، الذي، هي، هو، مع، عن، أن، أو، ... in, from, on, to, which, she, he, with, about, that, or, ...
D_1: Religion and belief systems	الله، كتاب، أب، عبد، إسلام، مسيحي، قرن، قدس، كنيسة، محمد، عهد، ... god, book, august, slave, Islam, Christian, a century, holy, a church, Mohammed, covenant, ...
D_3: History and events	معركة، قوة، عسكري، حرب، بريطانيا، ألماني، قرن، قائد، جبهة، جيش، ... battle, energy, soldier, war, Britain, German, a century, leader, front, army, ...
D_5: Culture and the arts	فيلم، لغة، مسلسل، إخراج، ممثل، موسيقى، كتاب، رواية، رسم، ... movie, language, a series, directed by, actor, music, book, a story, draw, ...
D_7: Mathematics and logic	رياضي، هندسة، إقليدي، مسافة، نظري، تفاضلي، جبري، لوغاريتم، تساوي، ... mathematician, geometry, euclidean, distance, theoretical, differential, algebraic, logarithm, equal, ..

Table VII shows the evaluation metrics results for all domains using our technique and the entropy. Resulting values confirm that our technique exceeds the entropy either for domain-independent stop-words (D_0) or domain-dependent stop-words (D_1 to D_{10}) detection.

Table VIII bellow gives examples of the obtained aggregate list.

Finally, this experiment allows us to approve that the proposed model uses frequency and distribution metrics to detect domain-independent and domain-dependent stop-words. However, we cannot conclude that our model can be applied on any corpus. To ensure that, we re-evaluate our technique using a second corpus.

A. *Results confirmation*

To confirm our automated stop-words detection model, we re-evaluate it using another corpus by gathering articles in different domains from an electronic newspaper site. Table IX describes the collected articles and the obtained stop-words lists.

Table IX: Newspaper corpus statistics and results

Domain	Documents number	Words number	Unique words number	Detected stop-words number
Art and culture	389	178340	29725	62
economy	389	138308	20186	74
history	115	271874	17031	39
international	346	123361	23502	50
Interviews	119	256677	25892	49
Media	271	242220	33214	43
Moroccans of the world	309	106808	22739	49
opinions	278	288969	47242	59
Orbits	301	172966	30690	75
Policy	1560	559804	23107	87
regions	511	220923	27616	56
Sport	110	32999	6524	39
Tamazight	224	130736	20674	48
All domains randomly	1500	621496	57993	60

Analyzing the results obtained from table IX, we discover that the number of detected stop-words vary when we use different corpora. For example, for the newspaper corpus domain “Art and culture” we detect 62 stop-words, while for the Wikipedia corpus domain “Culture and the arts” we detect 34 stop-words. This is a conflicting result that remind us the use of the “own corpora” trouble previously mentioned. However, after a careful analysis, we observe that the number of documents can be the factor that alters the detection system (389 documents for newspaper corpus “Art and culture” domain versus 130 documents for Wikipedia corpus “Culture and the arts” domain).

Table X: Detection results for domain-independent stop-words

Documents number	Stop-words number		Common stop-words	
	Newspaper corpus	Wikipedia corpus	Number	Percentage
30	14	12	10	83.33
70	23	21	18	85.71
150	35	30	23	76.67
350	42	38	28	73.68
500	47	51	40	85.10

To check this hypothesis, we apply our technique to different sets of documents from the two corpora for both domain-independent and domain-independent documents. Table X below summarizes the obtained results for the domain-independent documents.

As we can see, by applying our technique on the first set composed with 30 documents, we obtain 14 stop-words using the newspaper corpus, 12 stop-words using the Wikipedia corpus and 10 stop-words in common between the two lists which accounts for 83.33% of the detected stop-words.

The analysis of the obtained results demonstrates that the minimum rate of the common stop-words is 73.68% obtained with a set of 350 documents. These domain-independent stop-words results confirm our hypothesis saying that the number of documents is the factor that alters the detection system.

To confirm our hypothesis, we hold the same evaluation on the domain-dependent stop-words by detecting the stop-words of the “Art and culture” domain. Assessing the obtained results in Table XI, domain-dependent stop-words hold the particular feature of having not only common similar stop-words but also common equivalent stop-words. A common similar is the same stop-word which appears in both lists, whereas, common equivalent stop-words are two stop-words with different lexemes but having the same meaning such as “writer and author”.

Table XI: Detection results for “Art and culture” domain

Documents number	Stop-words number		Common stop-words	
	Newspaper corpus	Wikipedia corpus	Number (Similar + Equivalent)	Percentage
20	19	18	11 (7 + 4)	61.11
50	23	21	14 (9 + 5)	66.67
75	27	25	20 (11 + 9)	80.00
100	30	28	23 (12 + 11)	82.14
130	35	34	27 (14 + 13)	79.41

For instance, for the set composed with 50 documents, we obtain 23 stop-words using the newspaper corpus and 21 stop-words using the Wikipedia corpus. In this set, we have 9 common similar stop-words appearing in both lists such as “ثقافة (culture), سينما (cinema), شاعر (poet), مسرح (theatre)” and 5 common equivalent stop-words such as “شاشة (screen) against تلفزيون (TV), كاتب (writer) against مؤلف (author) or صورة (picture) againstتصوير (photography)”.

The number of common stop-words exceeds 60% in all sets and is around 80% in the sets composed of more than 75 documents. This rate in relatively low for the sets with a few number of documents but this can be explained by the nature of the documents in the two corpora. Because newspaper corpus contains documents related to current events, while Wikipedia corpus deals with documents for all intents and purposes such as the history of the arts, the visual arts, the literary arts and the performing arts.

In this way, we can conclude that a low number of documents is the factor that can alter our detection system for both domain-dependent and domain-independent stop-words.

Finally, this evaluation lets us to affirm that the proposed model detects domain-independent and domain-dependent stop-words using frequency and distribution metrics, regardless of the used corpus.

VI. Conclusion

Stop-words detection in existing works is based generally on generic lists or using some frequency based metrics. However, this detection process depends on the documents and the corpus being used for their detection. Consequently, there is a mass of stop-words lists with a variable content. In this work, we conceive a novel method that detects not only domain-independent stop-words but also domain-dependent stop-words, regardless of whether the corpus involves a specific domain or not. We adopt a bi-technical model for stop-words detection. The first one is a frequency-based one and provides us the first stop-words list. Whereas the second one is a distance-based process using a customized vector space model carrying the second stop-words list. The aggregation of the two lists leads to the final stop-words list. The developed method is experimented and evaluated using two different corpora reaching an average detection rate of 91.85% for the F-measure metric.

VII. References

[1] AL-SHALABI, Riyadh, KANAAN, Ghasan, JAAM, Jihad M., et al. Stop-word removal algorithm for Arabic language. In: Proceedings of 1st International Conference on Information & Communication Technologies: from Theory to Applications, CTTA'04. 2004. p. 545-550.

[2] ABU EL-KHAIR, Ibrahim. Effects of stop words elimination for Arabic information retrieval: a comparative study. International Journal of Computing & Information Sciences, 2006, vol. 4, no 3, p. 119-133.

[3] ZOU, Feng, WANG, Fu Lee, DENG, Xiaotie, et al. Automatic construction of Chinese stop word list. In: Proceedings of the 5th WSEAS international conference on Applied computer science. 2006. p. 1010-1015.

[4] SAVOY, Jacques. A stemming procedure and stopword list for general French corpora. JASIS, 1999, vol. 50, no 10, p. 944-952.

[5] ZHENG, Gong et GAOWA, Guan. The selection of Mongolian stop words. In: Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on. IEEE, 2010. p. 71-74.

[6] MEDHAT, Walaa, YOUSEF, Ahmed H., et KORASHY, Hoda. Corpora Preparation and Stopword List Generation for Arabic data in Social Network. arXiv preprint arXiv:1410.1135, 2014.

[7] DAVARPANAH, Mohammad Reza, SANJI, M., et ARAMIDEH, M. Farsi lexical analysis and stop word list. Library Hi Tech, 2009, vol. 27, no 3, p. 435-449.

[8] KUMARAN, Giridhar et ALLAN, James. Text classification and named entities for new event detection. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004. p. 297-304.

[9] SANGAIAH, Arun Kumar, FAKHRY, Ahmed E., ABDEL-BASSET, Mohamed, et al. Arabic text clustering using improved clustering algorithms with dimensionality reduction. Cluster Computing, 2018, p. 1-15.

[10] Glossbrenner, Alfred, and Emily Glossbrenner. Search engines for the world wide web. Peachpit Press, USA, 2001.

[11] METIN, Senem Kumova et KARAOĞLAN, Bahar. STOP WORD DETECTION AS A BINARY CLASSIFICATION PROBLEM. Anadolu University Journal of Science and Technology A-Applied Sciences and Engineering, 2017, vol. 18, no 2, p. 346-359.

[12] HAO, Lili et HAO, Lizhu. Automatic identification of stop words in chinese text classification. In: Computer Science and Software Engineering, 2008 International Conference on. IEEE, 2008. p. 718-722.

[13] CHEKIMA, Khalifa et ALFRED, Rayner. An Automatic Construction of Malay Stop Words Based on Aggregation Method. In: International Conference on Soft Computing in Data Science. Springer, Singapore, 2016. p. 180-189.

[14] ALAJMI, A., SAAD, E. M., et DARWISH, R. R. Toward an ARABIC stop-words list generation. International Journal of Computer Applications, 2012, vol. 46, no 8, p. 8-13.

[15] SALTON, Gerard, WONG, Anita, et YANG, Chung-Shu. A vector space model for automatic indexing. Communications of the ACM, 1975, vol. 18, no 11, p. 613-620.